

An Empirical Assessment of the Properties of Inverted Generational Distance on Multi- and Many-Objective Optimization

Leonardo C. T. Bezerra^{1,2}, Manuel López-Ibáñez³, and Thomas Stützle⁴

¹ DCC-CI, Universidade Federal da Paraíba (UFPB), João Pessoa, PB, Brazil
`leo.tbezerra@ci.ufpb.br`

² Alliance Manchester Business School, University of Manchester, UK
`manuel.lopez-ibanez@manchester.ac.uk`

³ IRIDIA, Université Libre de Bruxelles (ULB), Brussels, Belgium
`stuetzle@ulb.ac.be`

Abstract. The inverted generational distance (*IGD*) is a metric for assessing the quality of approximations to the Pareto front obtained by multi-objective optimization algorithms. The *IGD* has become the most commonly used metric in the context of many-objective problems, i.e., those with more than three objectives. The averaged Hausdorff distance and IGD^+ are variants of the *IGD* proposed in order to overcome its major drawbacks. In particular, the *IGD* is not Pareto compliant and its conclusions may strongly change depending on the size of the reference front. It is also well-known that different metrics assign more importance to various desired features of approximation fronts, and thus, they may disagree when ranking them. However, the precise behavior of the *IGD* variants is not well-understood yet. In particular, IGD^+ , the only *IGD* variant that is weakly Pareto-compliant, has received significantly less attention. This paper presents an empirical analysis of the *IGD* variants. Our experiments evaluate how these metrics are affected by the most important factors that intuitively describe the quality of approximation fronts, namely, spread, distribution and convergence. The results presented here already reveal interesting insights. For example, we conclude that, in order to achieve small *IGD* or IGD^+ values, the approximation front size should match the reference front size.

Keywords: multi-objective optimization, performance assessment, inverted generational distance

1 Introduction

Due to the conflicting nature of the multiple objectives to be optimized in a single run, the goal of an EMO algorithm is to find a set of high-quality, trade-off solutions, rather than a single one. Such solution sets are approximations to the Pareto-optimal front. Pareto-optimality only provides a partial ranking between such approximation fronts. Thus, their relative quality is typically evaluated with the aid of *quality metrics*, also known as quality indicators [14], which

provide a complete ranking. Many quality metrics have been proposed in the literature [9, 11–14], and multiple quality metrics are often used simultaneously, because each metric assigns different importance to various desirable features of approximation fronts, such as convergence, spread and distribution. These differences between metrics may lead to “disagreements”, when each metric chooses different approximation fronts as the best ones. Empirical studies have shown that the degree of disagreement strongly depends on features of the approximation fronts, such as convexity, and on the correlation between objectives and their number [10]. Understanding the properties of these metrics is critical for correctly selecting which metrics to use and interpreting their outcome.

One of the most desirable properties of a quality metric is Pareto compliance. A quality metric is Pareto-compliant if, and only if, the ranking it establishes over approximation fronts does not contradict Pareto optimality [14]. In other words, it cannot happen that the metric ranks one front better than another while the latter would always be preferred according only to Pareto optimality. The use of a non-Pareto-compliant metric to evaluate algorithms that attempt to approximate the Pareto front may lead an analyst to prefer an algorithm that returns approximation fronts that are strictly worse in terms of Pareto optimality.

Three widely-used unary quality metrics are the hypervolume (I_H [13]), the (additive or multiplicative) epsilon (I_ϵ^1 [14]) and the *inverted generational distance* (IGD [4]). Both I_H and I_ϵ^1 are Pareto-compliant [14],⁴ whereas the IGD is not [9]. Despite this drawback, the IGD has become widely adopted in EMO studies of many-objective optimization problems, i.e., problems with more than three objectives. A possible explanation is that the IGD is cheaper to compute than I_H , since the computational cost of I_H grows exponentially with the number of objectives [1]. Moreover, it is commonly assumed that both IGD and I_H are able to measure the desired features of approximation sets, that is, convergence, spread and distribution. However, recent empirical studies have shown that the disagreement between I_H and IGD increases with the number of objectives [2, 10]. Thus, understanding what is exactly being measured by the IGD under various scenarios is of critical importance.

Alternative versions of the IGD have been proposed in recent years [9, 11] with hopes of addressing its potential drawbacks. In addition to its lack of Pareto-compliance, the IGD is not strictly a distance since under some conditions it does not satisfy the triangle inequality among approximation fronts [11]. Moreover, the IGD may evaluate two approximation fronts as almost equal if the main difference between the two is that one of them contains a very poor objective vector (an outlier). Another potential drawback of the IGD is that its value is quite sensitive to changes in the size of the reference front, that is, a finer-grained reference front may significantly alter conclusions previously obtained with a smaller reference front, even if both reference fronts only present optimal solutions. The averaged Hausdorff distance (Δ_p [11]) is a variant of IGD that attempts to alleviate some of these drawbacks, except for the lack of Pareto-compliance. More recently, the IGD^+ has been proposed as a Pareto-compliant

⁴ To be more precise, the ϵ -metric is only weakly Pareto-compliant, but we do not make a distinction between weakly and non-weakly Pareto-compliance in the remainder of this paper.

variant of IGD [9], being very similar to the IGD and being as robust as Δ_p to different sizes of reference fronts. The properties of the IGD^+ are still poorly understood and so far no investigation has been conducted about the behavior of the IGD^+ with respect to convergence, spread, and distribution.

In this work, we conduct an empirical investigation specifically targeting IGD and its Pareto-compliant variant IGD^+ . Concretely, we generalize an existing bi-objective benchmarking problem [7] for any number of objectives and we consider scenarios with up to ten objectives. Then, we design a series of experiments where approximation fronts are evolved for increasing convergence, spread, or distribution, and evaluate how the selected metrics respond to these changes. Effectively, our experiments isolate the effects of convergence, spread, and distribution. Furthermore, we design two experiments that simulate practical scenarios that EMO algorithms may run into, namely, when the approximation front has converged to the central region of the Pareto front and when the approximation front has achieved a good convergence and maximum spread but its distribution is poor.

Our results show that the IGD and IGD^+ behave exactly the same for all the practical purposes considered in this work. In addition, the factors that affect one variant affect the other in the same degree. For instance, we observe that the most important feature to ensure low IGD values is to have an approximation front that matches the size of the reference front adopted. Knowledge of this feature is critical as the default practice in the performance assessment of EMO algorithms is to use very large reference fronts, and here we demonstrate that in this circumstance the IGD and IGD^+ values may start worsening even if the spread of the approximation front is improving without worsening any other desirable feature. Another important insight concerns the effect of the parameter meant to regulate the importance of outliers. We observed that a setting often used in the literature leads to a constant IGD value despite changes in the distribution of the approximation front.

The remainder of this work is structured as follows. In Section 2, we briefly review the most relevant conceptual definitions related to the performance assessment of EMO algorithms, highlighting the desirable features of approximation fronts that we use as factors in our empirical investigation. Next, Section 3 presents an overview of IGD variants and explains why we focus on IGD and IGD^+ . Section 4 details our experimental setup, and Section 5 reports the empirical investigation we conduct. Finally, Section 6 presents our conclusions and discussion of future work.

2 Performance assessment of EMO algorithms

In multi-objective optimization (MO), the goal is to simultaneously optimize M objective functions.⁵ Therefore, the image of each potential solution is an *objective vector* with M components. The conflicting nature of objectives typically prevents the existence of a single, globally optimal solution that optimizes all objectives at once. In the absence of preference information regarding the

⁵ In the following we assume maximization, without loss of generality.

importance of each objective, solutions are often compared in terms of Pareto-optimality, where a solution with objective vector a is said to *dominate* another solution with objective vector b iff $a_i \geq b_i, \forall i = 1, \dots, M$ and $\exists j a_j > b_j$. Two objective vectors are mutually nondominated if none of them dominates the other. The goal then becomes to find the set of Pareto-optimal solutions, that is, those solutions that are not dominated by any other feasible solutions; or rather the image of this set in the objective space, the Pareto front.

Since finding the Pareto front is often intractable, EMO algorithms attempt to find a high-quality approximation of it, namely, an *approximation front* composed of mutually nondominated objective vectors. Hence, the performance assessment of EMO algorithms requires the evaluation of the relative quality of approximation fronts. Although Pareto-optimality may sometimes be enough to conclude that one approximation front is better than another, the most common case is that fronts are mutually incomparable. Nonetheless, there are features that, in addition to Pareto-optimality, are desirable in high-quality approximations [10]:

1. **Convergence** refers to the (near-)optimality of individual solutions. A front is said to have converged if all of its solutions are Pareto-optimal.
2. **Spread** refers to the extent of the front, more specifically to the distance between the extreme solutions of a front.
3. **Distribution** refers to the evenness of the front, more specifically to the uniformity of the distances between pairs of adjacent solutions.

Instead of directly measuring each individual feature, quality metrics can be found in the literature that attempt to evaluate all features at once [9, 11–14]. However, each metric assigns a different importance to each feature and behaves differently depending on the characteristics of the problem and the particular fronts being evaluated, thus it is common that multiple metrics are used for performance assessment. In particular, experiments have shown [10] that the *IGD* and the *I_H* consistently disagree in typical scenarios arising in many (more than three) objective problems. Hence, understanding the behavior of quality metrics under various scenarios is crucial for performance assessment. While the *I_H* and, to some extent, the *IGD* are fairly well understood nowadays, newer variants such as *IGD⁺* have received little attention.

3 The inverted generation distance and its variants

The predecessor of the *IGD*, the **generational distance** (*GD* [12]), was proposed nearly two decades ago. The *GD* is defined as the distance between each objective vector a in a given approximation front A and the closest objective vector r in a reference front R , which is either the actual Pareto front or a very good approximation to it, averaged over the size of A . Formally,

$$GD(A, R) = \frac{1}{|A|} \left(\sum_{a \in A} \min_{r \in R} d(a, r)^p \right)^{1/p}, \quad d(a, r) = \sqrt{\sum_{k=1}^M (a_k - r_k)^2} \quad (1)$$

A value of $p = 2$ was used in the original proposal, but this choice was later superseded by $p = 1$ for simplicity of interpretation and computation. With

$p = 1$, the *GD* becomes an average of the Euclidean distances between each objective vector in A and its closest objective vector in R . The *GD* metric is fast to compute and correlates with convergence to the reference set. However, the *GD* is not Pareto-compliant [14] and it is also sensitive to the size of the approximation front A . Thus, large approximation fronts of poor quality may be ranked highly by *GD*.

The **inverted generational distance** (*IGD* [4]) was proposed as an improvement over the *GD* based on the very simple idea of reversing the order of the fronts considered as input by the *GD*, i.e., $IGD(A, R) = GD(R, A)$. In other words, the *IGD* equals the *GD* metric but computing the distance between each objective vector in the reference front and its closest objective vector in the approximation front, averaged over the size of the reference front. Parameter p plays a similar role as in the *GD* and often defaults to $p = 1$. The *IGD* is not sensitive to the size of the approximation fronts and it provides a ranking that intuitively matches more closely the desirable convergence, spread and distribution. Since it is also computationally fast to compute, *IGD* soon became the most widely used metric to assess many-objective EMO algorithms. Nonetheless, the *IGD* has been shown recently to lack Pareto-compliance [9].

The **averaged Hausdorff distance** (Δ_p [11]) was proposed as an attempt to address three potential drawbacks of the *IGD*. First, despite being characterized as a distance metric, the *IGD* sometimes violates the triangle inequality property. Second, the size of the reference front has a significant effect on the *IGD* values, to the point that adding additional solutions to the current reference set may change the relative ranking of approximation fronts. Third, if the main difference between two approximation fronts is that one contains a clearly poorer objective vector, the *IGD* may still regard both fronts as roughly equal, in other words, the *IGD* is often lenient about outliers. To overcome the first and third drawbacks, Δ_p is defined as an averaged Hausdorff distance metric, regulated by the numerical parameter p . In particular, larger values of p mean stronger penalties for outliers and fewer triangle inequality violations. Concerning the second drawback, Δ_p uses an alternative version of *IGD* (IGD_p), where the denominator is also affected by the parameter p . The formal definition of Δ_p is given below:

$$\Delta_p(A, R) = \max(IGD_p(A, R), IGD_p(R, A)) \quad (2)$$

$$IGD_p(A, R) = \left(\frac{1}{|R|} \sum_{r \in R} \min_{a \in A} d(r, a)^p \right)^{1/p} \quad (3)$$

Finally, the **modified inverted generational distance** (IGD^+ [9]) proposes the following modification of the distance function of the original *IGD*. For objective vectors that are dominated by the reference front, the traditional Euclidean distance is adopted. However, for objective vectors that are nondominated w.r.t. to the reference front, only the dominated objective vector components are used for computing the distance. Formally, given a problem where all M objectives must be maximized, the distance function d in Eq. (3) is replaced

by:

$$d^+(r, a) = \sqrt{\sum_{k=1}^M (\max\{r_k - a_k, 0\})^2} \quad (4)$$

This modification is enough to make IGD^+ weakly Pareto compliant, similarly to I_ϵ^1 . In addition, the definition of IGD^+ includes the denominator $|R|$ under the exponent $1/p$, as in the IGD_p definition proposed for the Δ_p (Eq. 3).

Nowadays, the IGD has effectively superseded GD , however, Δ_p has not gained the attention of the community as a widely used performance assessment metric. A possible explanation is that Δ_p still lacks Pareto-compliance and the drawbacks addressed by Δ_p only arise in unusual scenarios, such as for very small reference fronts [11]. On the other hand, IGD^+ is a small modification of IGD that adds Pareto-compliance, thus IGD^+ seems a more likely candidate to supersede IGD . Yet, there is little understanding so far about the behavior of IGD^+ [8], with current investigations focused on its Pareto-compliance. In the remainder of this paper, we experimentally compare the behavior of IGD and IGD^+ under various scenarios.

4 Experimental setup

In the following, we conduct a series of experiments in order to understand how IGD and IGD^+ react to various desirable features of approximation fronts. A first set of experiments is designed to evaluate each feature in isolation. In addition, we design experiments that simulate common scenarios in the context of EMO, where features are not isolated.

Benchmark problem. As a starting point, we consider the bi-objective optimization problem designed by Ishibushi et al. [7] that presents a linear-shaped Pareto front (Fig. 1, left). Generalized to any number of objectives M (Fig. 1, right), we have that a solution s is optimal if $\sum_{i=1}^M f_i(s) = 10$, where $f_i(s)$ is the i -th objective value of solution s . In this work, we study $M \in \{2, 3, 5, 10\}$.

Reference fronts. Since generating optimal solutions is trivial, we produce reference fronts of different resolution by using different front sizes. In the case of $M \in \{2, 3\}$, each reference front $R_{\langle M, d \rangle}$ is created using a uniform weight vector generation method, parameterized by the number of divisions d , and its size equals $\binom{M-1+d}{d}$. In the case of $M \in \{5, 10\}$, a uniform distribution of weights generates fronts with many more solutions on the extremes of the objective space than in the center. Instead, we adopt the two-layer approach proposed by [5]. In more detail, for a given d value, we generate a uniform set of $\lfloor d/2 \rfloor$ weights on the extremes of the objective space where are least one objective function is equal to zero (the outer layer), and a uniform set of $d - \lfloor d/2 \rfloor$ weights in the center of the objective space (the inner layer). We consider three different d values, representing small, moderate and large reference fronts, for each value of M , as shown in Table 1.

Approximation fronts. In general, we generate approximation fronts using the same method explained above for reference fronts, that is, according to the number of objectives M and a parameter d (Table 1).

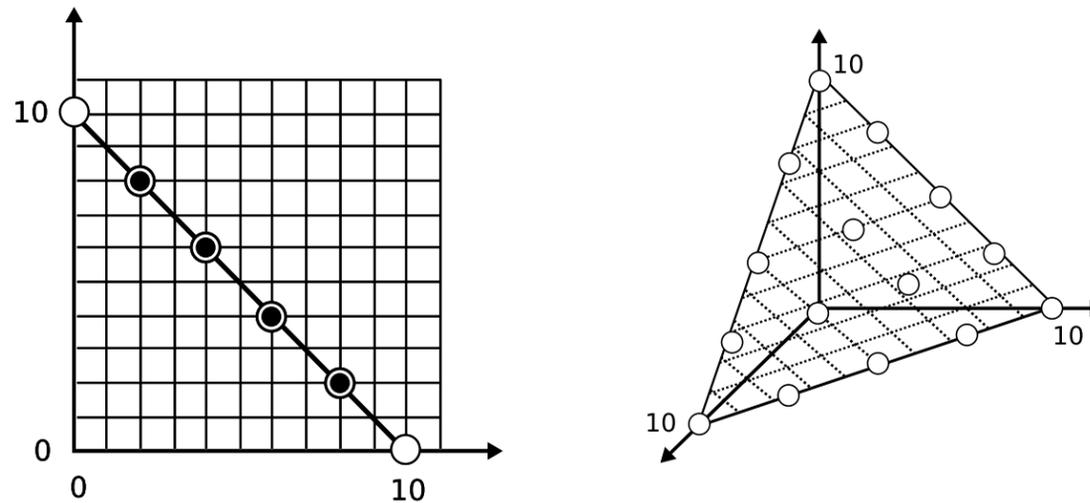


Fig. 1. Linear-shaped maximization problem suggested by [7]. Black dots represent approximation front solutions, whereas white circles represent reference front solutions created with $M = 2$, $d = 3$ (left) and $M = 3$, $d = 5$ (right).

Table 1. Size of the generated reference front ($|R|$) for each number of objectives (M) and value of parameter d . We use the same method to generate approximation fronts.

M	d	$ R $									
2	3	4	3	6	28	5	5	50	10	4	110
	19	20		13	105		8	140		5	275
	99	100		19	210		13	540		7	935

Effect of p . As explained earlier, most of the literature only considers $p = 1$ while some works use $p = 2$. To understand how different values of p affect the behavior of the quality metrics, we consider $p \in \{1, 2, 3\}$.

Seeds for random number generation. When sampling is adopted, we repeat each experiment 25 times with a common set of random seeds in order to reduce variance between experiments.

5 Empirical assessment of IGD and IGD^+

5.1 Desirable features of approximation fronts

We first explain how we designed experiments to evaluate one desirable feature of approximation fronts at a time. After that, we discuss our conclusions from these experiments.

Convergence. To evaluate convergence, we consider approximation fronts that are obtained by translating a reference front until it intersects with the axes. These approximation fronts are then iteratively “evolved” using a linear interpolation for each pair $\langle a', r' \rangle$, where a' is the solution from the approximation

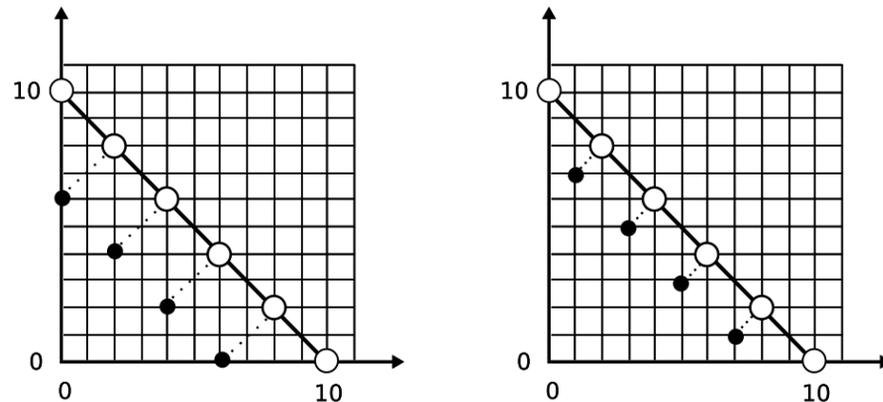


Fig. 2. Illustration of the experiments where convergence is isolated. Left: approximation front in its initial state after translation. Right: approximation front after 50 iterations of linear interpolation. All intermediate states present the same spread and distribution.

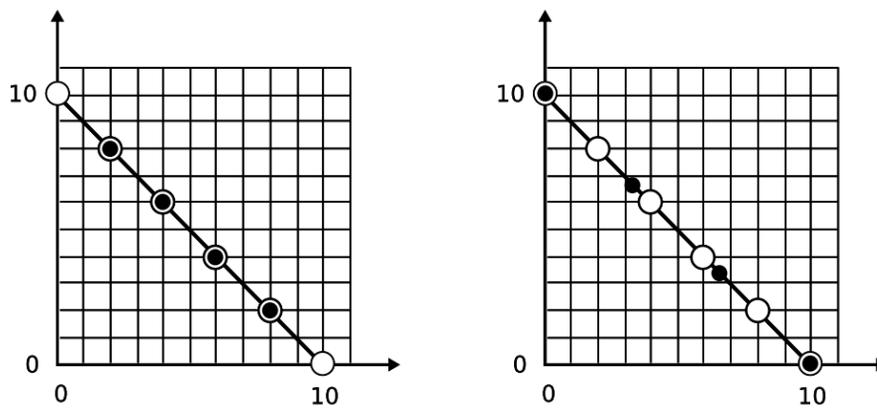


Fig. 3. Illustration of the experiments where spread is isolated. Approximation front with a spread of 6 (left) and after 100 iterations of evolution, with a spread of 10 (right). All iterations present the same distribution and convergence.

front obtained by translating reference front solution r' . To measure how far the approximation front is from the reference front at a given iteration, we compute the Euclidean distance for the pair $\langle a', r' \rangle$. Each iteration reduces this distance by the same value until it becomes zero, and the step value is calculated in order to perform 100 iterations. Figure 2 illustrates a translated front before being evolved (left) and at a later iteration (right).

Spread. We evaluate spread by initially selecting the desired distance along a single objective between the extreme solutions. Next, we multiply this distance by the uniform weight set obtained for the given M and d , and translate this front so that its solutions become optimal. By increasing the distance between the extreme solutions, we are able to iteratively generate approximation fronts with increasing spread but the same convergence (since all solutions are optimal) and distribution (since solutions are equally distributed between the extremes). Figure 3 illustrates an approximation front for $M = 2$ with spread of 6 (left) and with a spread of 10 (right).

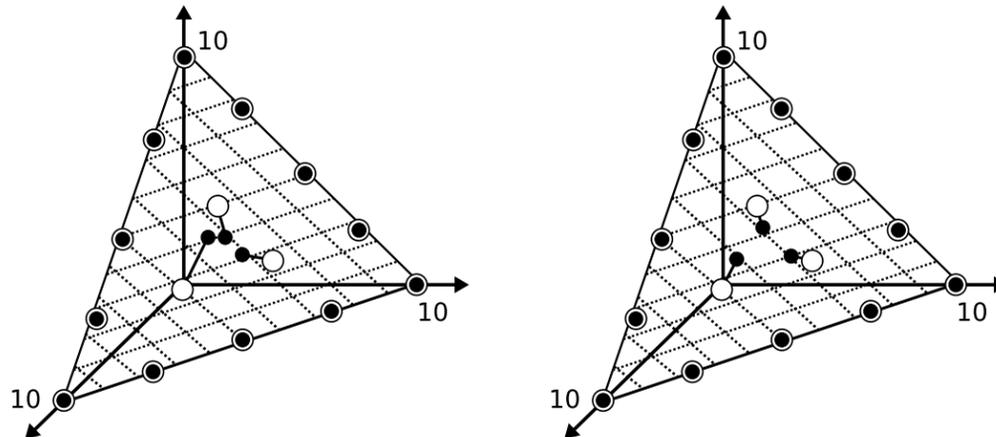


Fig. 4. Illustration of the experiments where distribution is isolated. Left: approximation front with maximum spread and inner layer solutions sampled around the center of the objective space with a Gaussian distribution. Right: intermediate stage of the evolutionary process, where each solution is getting closer to its point of destination.

Distribution. To isolate distribution, we generate approximation fronts in two steps. First, we copy the outer layer (those objective vectors that have at least one zero component) from the reference front of the same size. Next, we sample n_{inner} Pareto-optimal solutions by sampling coordinates using a Gaussian distribution around the center of the Pareto front, where $n_{\text{inner}} = n - n_{\text{outer}}$, n is the size of the corresponding reference front, and n_{outer} is the amount of solutions in the outer layer. To sample solutions that are concentrated around the center of the objective space, we use the algorithm provided in the supplementary material [3]. Figure 4 (left) illustrates a front generated with this method. Once a front is sampled, we evolve it towards an even distribution. More precisely, we first associate each approximation front solution a^r with the point r that is closest to it (in terms of Euclidean distance) in the reference front of the same size. Next, at each iteration, we use linear interpolation to translate each a^r towards its corresponding r , so that the resulting approximation front is more evenly distributed than the previous one. Figure 4 (right) illustrates an intermediary stage of evolution of the front depicted in Figure 4 (left). Since only solutions from the inner layer are sampled, only those solutions are translated at each intermediary stage. In addition, both convergence and spread remain constant, as all solutions are always optimal and the extreme solutions do not change. To compute how far a front is from a perfect distribution at a given iteration, we define an entropy-like metric defined as $e(A, R) = \frac{1}{|R|} \sum_{r \in R} d(a^r, r)$. The difference between this e metric and the IGD with $p = 1$ is that e does not allow a reference front solution to be associated with multiple approximation front solutions. Effectively, this metric can be seen as a simplification of the *root mean square error*, an average of how far each solution a_i is from where it should be.

We next discuss the most important, high-level insights we observe in the results produced.

Effect of front sizes. In the case of varying spread and distribution, we observe that, whichever M considered, the best IGD and IGD^+ values are obtained when

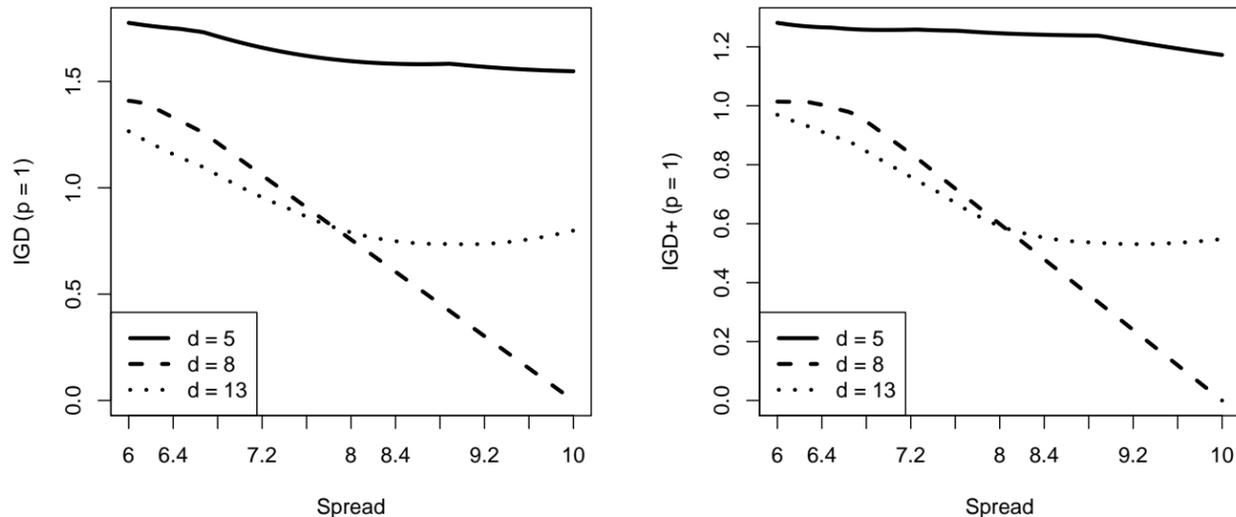


Fig. 5. Effect on IGD (left) and IGD^+ (right) of various approximation front sizes for increasing spread. Reference front is generated with settings $M = 5$, $d = 8$. Approximation fronts are generated with settings $M = 5$ and $d \in \{5, 8, 13\}$. Notice that, despite different in ranges, the shapes of the curves for both IGD and IGD^+ are similar.

the approximation front size matches the reference front size. That is, the IGD metrics will rank an approximation front of the same size as the reference front better than a larger or smaller front, and the difference will increase with higher spread. This result is specially counter-intuitive in the case of fronts larger than the reference front, as one would expect that, everything else being equal, larger approximation fronts are to be preferred. This is illustrated in Figure 5, depicting the response of IGD and IGD^+ when $M = 5$, in dependence of different spread values (x -axis) and different approximation set sizes ($d \in \{5, 8, 13\}$), using a fixed reference set size ($d = 8$). Both IGD and IGD^+ metrics assign the best quality to the largest size approximation front when spread is small. However, when the spread of the front gets close to the maximum tested, only the quality of the front with the same size as the given reference front ($d = 8$) continues to improve until the minimal possible IGD value is obtained. For the largest approximation front size ($d = 13$), the quality according to IGD and IGD^+ even worsens again, despite neither the convergence or distribution are actually worse.

Similarity between IGD and IGD^+ results. The main difference between IGD and IGD^+ is that only components of objective vectors dominated by the reference front contribute to the computation of the distance. This means that, as observed, the IGD and IGD^+ produce equal results for approximation fronts that are dominated by the reference front, however, one would expect strong differences when the approximation front is mostly nondominated with respect to the reference front, as is the case in most of our experiments. Our experiments show that, under the same conditions, the IGD and IGD^+ differ in range but their behavior with respect to changes in convergence, spread and distribution is very similar. This is observed by comparing the shapes of the curves corresponding to IGD and IGD^+ in most of our plots, for example, in Figure 5. The only

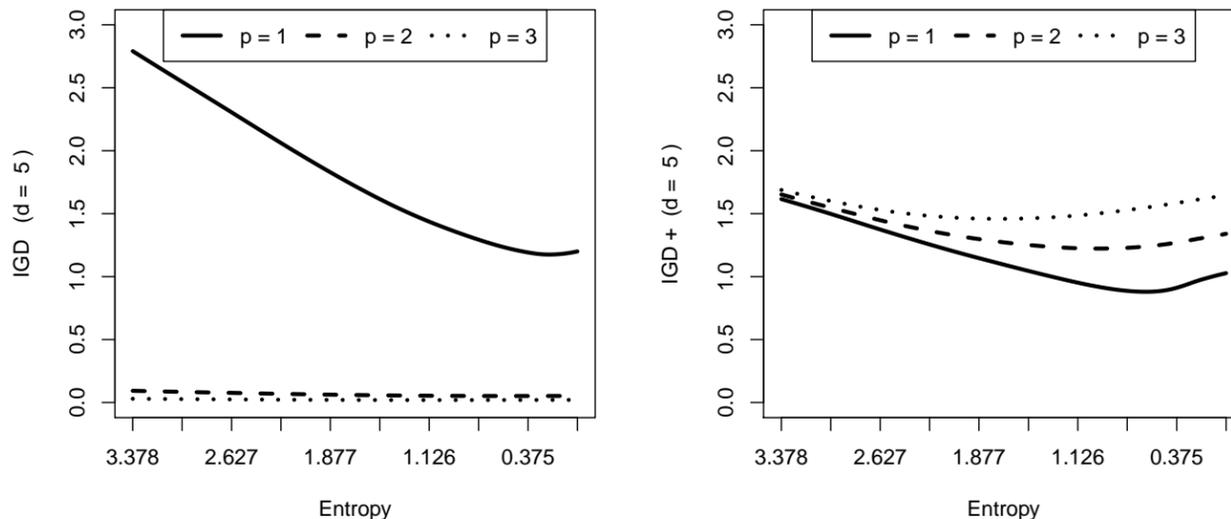


Fig. 6. Illustration of the effects of p . Distribution analysis of IGD (left) and IGD^+ (right) for $M = 10$, with an approximation front generated with $d = 5$ and a reference front generated with $d = 7$.

exception that we observed concerns changes in distribution when $M \in \{5, 10\}$ and the sizes of the approximation and reference fronts are not equal. Figure 6 depicts such a situation, where the shape of the curves when $p \geq 2$ is much smoother for IGD than for IGD^+ .

Effect of p . Under the conditions tested in this work, the only observable effect of p is changing the ranges of the IGD variants. However, it is interesting to notice that, while the range of the IGD decreases with larger p , the opposite is observed for the IGD^+ . This effect is shown in Figure 6. Another important effect that is also depicted in this figure concerns the IGD metric only. In particular, a value of $p = 2$, as sometimes used in the literature, makes the IGD insensitive to changes in distribution. This is a potentially dangerous limitation, and we observe that this effect becomes ever stronger with the increase in reference front sizes.

Particularities of many-objective problems. The convergence analysis when $M = 10$ differs from the overall patterns we have so far discussed, as in this case having larger approximation fronts is indeed a winning strategy. More precisely, when the reference fronts created with $d = 4$ or $d = 5$ are used, the approximation front created with $d = 7$ is either the best-performing according to both metrics, or at least very competitive. This situation is illustrated in Figure 7 for the IGD^+ (the same is observed for the IGD), and may be due to the ratio between solutions in the inner and outer layers. Concretely, when $d \in \{5, 7\}$, the number of solutions in the inner layer is far greater than the number in the outer layer, whereas for $d = 4$ this ratio equals one. Another hypothesis is that the IGD values become better as long as the approximation front does not have more points in the outer layer than the reference front. Understanding this behavior will require further analysis.

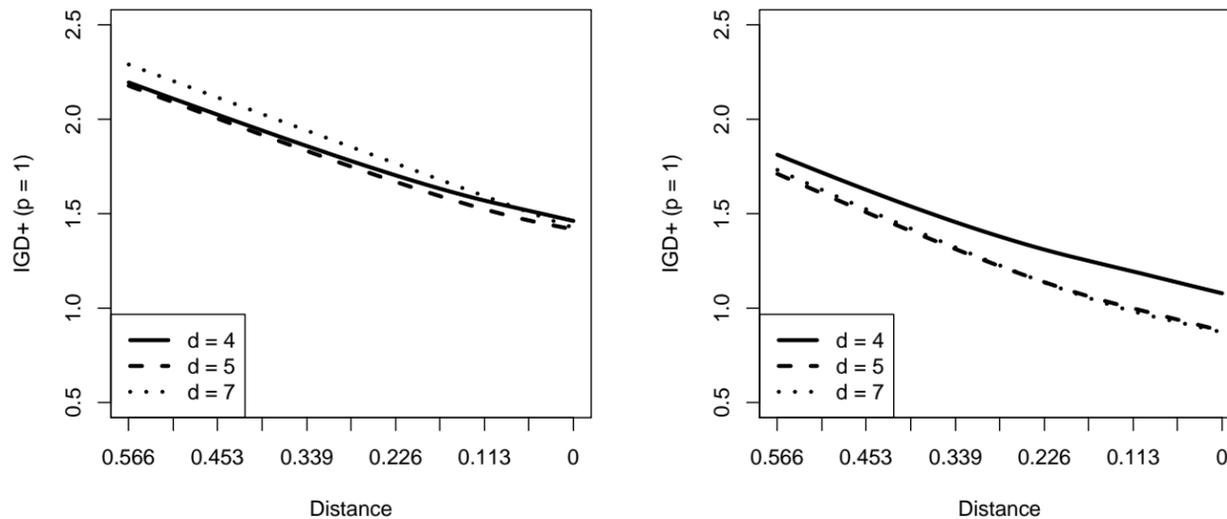


Fig. 7. Illustration of the particularities of many-objective optimization. IGD^+ convergence analysis indicate that, when $M = 10$, having larger approximation sets lead to better results. This effect is observed both when the reference front is created with $d = 4$ (left) and $d = 5$ (right).

5.2 Practical EMO scenarios

In addition to the experiments above, where we analyzed in isolation the desirable features of approximation fronts, we now design experiments that resemble the evolution of approximation fronts by an EMO algorithm. The first experiment simulates the scenario where an EMO algorithm may converge to the optimal front with maximum spread, but still needs to further improve the distribution of its approximation front. The second experiment simulates a scenario where an EMO algorithm reaches the optimal front but lacks both spread and distribution.

Distribution. In this scenario, the approximation front already has maximum spread and has converged (all of its solutions are Pareto-optimal), but still needs to improve distribution. This happens in practice when high-quality solutions can be found by decomposing the problem into single objective ones. This is the case when using scalarization-based local search to tackle the bi-objective permutation flowshop problem [6]. It may also happen when the EMO algorithm internally uses a quality metric that favors spread over distribution. We generate approximation fronts for this scenario by copying solutions from the corresponding reference front and then applying a small, random uniform perturbation to each non-extreme solution (inner layer solutions). Figure 8 (left) illustrates a front generated in this fashion when $M = 3$ and $d = 5$. We then simulate the evolution of the approximation front by translating the approximation front solutions towards their original position in the reference front by linear interpolation. The entropy-like metric e defined in the previous section is used here to measure the perturbation level of the intermediate fronts. Figure 9 (left) shows results for this scenario, where it can be seen that both IGD variants are

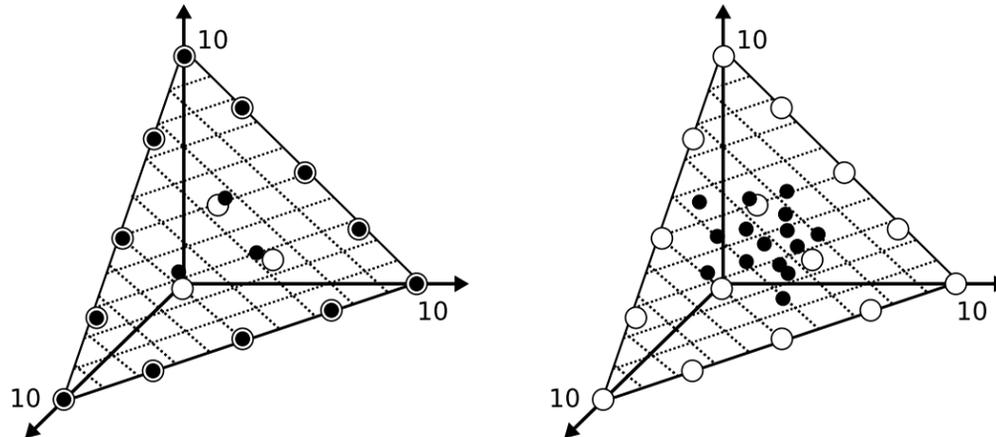


Fig. 8. Illustration of experiments where (left) inner layer solutions have been subject to a perturbation, and; (right) front has converged to the center of the objective space.

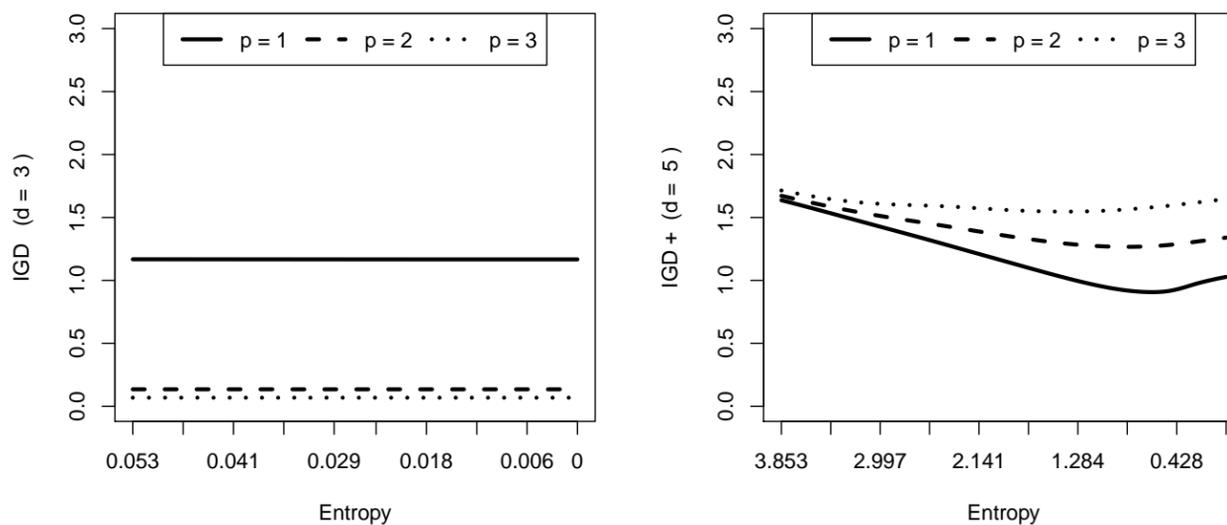


Fig. 9. Results for practical EMO scenarios. Left ($M = 2$): selected IGD variants are robust to perturbations in distribution for a small approximation set ($d = 3$) and a large reference set ($d = 99$). Right ($M = 10$): intermediate size approximation front ($d = 5$) gets worse IGD^+ values when trying to approximate a larger size reference front ($d = 7$), even though its distribution is improving.

barely affected by this kind of perturbation. When investigating the reason for such robustness, it becomes clear that this kind of perturbation introduces very little entropy to the approximation front, as evidenced in the x -axis of the plot depicted in Fig. 9 (left).

Distribution and spread. In this scenario, the algorithm has converged to the center of the objective space, lacking both spread and distribution. This scenario happens often in practice, whenever (i) the algorithm does not preserve extreme solutions, (ii) extreme solutions are hard to find, or (iii) the algorithm focuses

on converging to the Pareto front first, and spreading later. To simulate this scenario, we produce fronts using the same solution sampling used in the previous section to generate fronts with uneven distribution. We also use the entropy-like metric presented earlier to assess how far the front is from perfect distribution. The main difference with respect to those experiments is that, in this case, we do not generate an outer layer that maximizes the spread, but instead we sample additional objective vectors that are likely to be concentrated in the inner layer. Figure 8 (right) illustrates a front generated using this method. Results for this scenario confirm what was observed in the previous set of experiments, where the most important factor was matching the size of the approximation front with the size of the reference front. In fact, the striking similarity between Fig. 6 (right) and Fig. 9 (right) indicates that the outer layer kept to ensure maximum spread from the previous experiment did not play a relevant part in the response from the IGD metrics. This is an important finding, since it could explain the strong disagreements between the IGD and the I_H , that is, fronts with very good spread at the cost of distribution will be favored by I_H whereas IGD will likely do the opposite.

6 Conclusions

In this work, we have conducted a preliminary empirical investigation on the properties of two relevant quality metrics, namely the inverted generational distance (IGD [4]) and a variant that has been proven weakly Pareto-compliant (IGD^+ [9]). In particular, the IGD is perhaps the most adopted performance metric in the context of many-objective optimization due to its low computational cost and its ability to assess several desirable features of approximation fronts, namely, convergence, spread, and distribution. By contrast, IGD^+ is a recent variant that needs further investigation, a task we undertook in this work.

In order to analyze the behavior of the selected metrics, we designed a series of experiments that either (i) isolated the desirable features of approximation fronts, or (ii) simulated real-world situations often faced by EMO algorithms. Perhaps the most important insight is the observation that, under some circumstances, the value of IGD or IGD^+ may become worse if the size of the approximation front grows beyond the size of the reference front, even if convergence, spread and distribution remain equal. In addition, we have shown that, for the scenarios tested in this work, the selected metrics have a similar behavior. The only exception to this pattern concerns IGD , which we have shown to be unable to detect poor distribution depending on the value of the parameter p , which is sometimes not specified in the literature. Finally, another important observation concerns a possible explanation for the previously reported disagreement between I_H and IGD . We have shown that, under some circumstances, outer layer solutions have little impact on the value of the IGD .

Our investigation opens a number of possibilities for future analysis of quality metrics. The first and most straightforward step is to deepen the analysis of IGD variants, both by considering more variants (namely Δ_p) and more examples of problems with different geometries. A second step is to design yet more elaborate experiments to simulate other real EMO scenarios, besides the obvious

approach of assessing actual EMO algorithms in practice. Finally, it is imperative that these results be related to other metrics, specifically the hypervolume (I_H) and epsilon (I_ϵ^1) metrics, helping the community further understand their disagreements.

Acknowledgments. The research presented in this paper has received funding from the COMEX project (P7/36) within the IAP Programme of BelSPO. T. Stützle acknowledges support from the Belgian F.R.S.-FNRS, of which he is a senior research associate.

References

1. Beume, N., Fonseca, C.M., López-Ibáñez, M., Paquete, L., Vahrenhold, J.: On the complexity of computing the hypervolume indicator. *IEEE Trans. Evol. Comput.* 13(5), 1075–1082 (2009)
2. Bezerra, L.C.T.: A component-wise approach to multi-objective evolutionary algorithms: from flexible frameworks to automatic design. Ph.D. thesis, IRIDIA, École polytechnique, Université Libre de Bruxelles, Belgium (2016)
3. Bezerra, L.C.T., López-Ibáñez, M., Stützle, T.: An empirical assessment of the properties of inverted generational distance indicators on multi- and many-objective optimization: Supplementary material. <http://iridia.ulb.ac.be/supp/IridiaSupp2016-006/> (2016)
4. Coello Coello, C.A., Reyes-Sierra, M.: A study of the parallelization of a coevolutionary multi-objective evolutionary algorithm. In: Monroy, R., et al. (eds.) *Proceedings of MICAI, LNAI*, vol. 2972, pp. 688–697. Springer, Heidelberg, Germany (2004)
5. Deb, K., Jain, S.: An evolutionary many-objective optimization algorithm using reference-point-based nondominated sorting approach, part I: Solving problems with box constraints. *IEEE Trans. Evol. Comput.* 18(4), 577–601 (2014)
6. Dubois-Lacoste, J., López-Ibáñez, M., Stützle, T.: Improving the anytime behavior of two-phase local search. *Annals of Mathematics and Artificial Intelligence* 61(2), 125–154 (2011)
7. Ishibuchi, H., Akedo, N., Nojima, Y.: Behavior of multiobjective evolutionary algorithms on many-objective knapsack problems. *IEEE Trans. Evol. Comput.* 19(2), 264–283 (2015)
8. Ishibuchi, H., Masuda, H., Nojima, Y.: A study on performance evaluation ability of a modified inverted generational distance indicator. In: Silva, S., et al. (eds.) *GECCO*, pp. 695–702. ACM Press (2015)
9. Ishibuchi, H., Masuda, H., Tanigaki, Y., Nojima, Y.: Modified distance calculation in generational distance and inverted generational distance. In: Gaspar-Cunha, A., et al. (eds.) *EMO Part I, LNCS*, vol. 9018, pp. 110–125. Springer (2015)
10. Jiang, S., Ong, Y.S., Zhang, J., Feng, L.: Consistencies and contradictions of performance metrics in multiobjective optimization. *IEEE Trans. Cybern.* 44(12), 2391–2404 (2014)
11. Schütze, O., Esquivel, X., Lara, A., Coello Coello, C.A.: Using the averaged Hausdorff distance as a performance measure in evolutionary multiobjective optimization. *IEEE Trans. Evol. Comput.* 16(4), 504–522 (2012)
12. Van Veldhuizen, D.A., Lamont, G.B.: Multiobjective evolutionary algorithms: Analyzing the state-of-the-art. *Evol. Comput.* 8(2), 125–147 (2000)
13. Zitzler, E., Thiele, L.: Multiobjective evolutionary algorithms: A comparative case study and the strength Pareto evolutionary algorithm. *IEEE Trans. Evol. Comput.* 3(4), 257–271 (1999)

14. Zitzler, E., Thiele, L., Laumanns, M., Fonseca, C.M., Grunert da Fonseca, V.: Performance assessment of multiobjective optimizers: an analysis and review. *IEEE Trans. Evol. Comput.* 7(2), 117–132 (2003)